Parsed Page eXplorer (PPX): A Semantic Search Engine for Document Understanding

Levi Willms ● Ken Pu
Ontario Tech University ● Oshawa, ON, Canada
levi.willms@ontariotechu.net, ken.pu@ontariotechu.ca

1. Motivation

- **Problem**: Raw PDF documents lack machine-readable structure, making semantic queries impossible to spatially locate. Existing OCR tools provide bounding boxes but fragment text; LLMs reconstruct meaning but lose spatial information.
- **Challenge**: Aligning spatial OCR fragments with semantic LLM text to enable queries that return precise page coordinates.
- **Goal**: Establish semantic queries on PDF documents with verifiable source locations — establishing *semantic provenance* (tracing answers to their exact spatial origin).

2. Methodology

- The system will consist of two (2) workflows: **document parsing** and **document search**.
- **Document Parsing**: Analyze layout, generate markdown using PaddleOCR and LLM tools (MistralOCR, Gemini, OpenAI). [1, 3]
- **Document Search**: Align markdown with document layout to establish *semantic provenance*. Generate embeddings using sentence transformers. [2]
- **Validation**: Measure alignment accuracy and verify returned source locations are correct.
- **Evaluation**: Measure Q&A accuracy, localization precision, and alignment accuracy on ground truth dataset.

3. System

PPX (parse)

Layout analyzer

OCR (markdown)

MobileNetV2: Inverted Residuals and Linear Bottlenecks

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, Liang-Chieh Chen

Google Inc.

{sandler, howarda, menglong, azhmogin, lcchen}@google.com

Abstract

In this paper we describe a new mobile architecture, MobileNetV2, that improves the state of the art performance of mobile models on multiple tasks and benchmarks as well as across a spectrum of different model sizes. We also describe efficient ways of applying these mobile models to object detection in a novel framework we call SSDLite. Additionally, we demonstrate how to build mobile semantic segmentation models through a reduced form of DeepLabv3 which we call Mobile DeepLabv3.

The architecture is based on an inverted residual structure where the shortcut connections are between the thin bottleneck layers. The intermediate expansion layer uses lightweight depthwise convolutions to filter features as a source of non-linearity. Additionally, we find that it is important to remove non-linearities in the narrow layers in order to maintain representational power. We demonstrate that this improves performance and provide an intuition that led to this design.

Finally, our approach allows decoupling of the input/output domains from the expressiveness of the transformation, which provides a convenient framework for further analysis. We measure our performance on ImageNet [1] classification, COCO object detection [2], VOC image segmentation [3]. We evaluate the trade-offs between accuracy, and number of operations measured by multiply-adds (MAdd), as well as actual latency, and the number of parameters.

1. Introduction

Neural networks have revolutionized many areas of machine intelligence, enabling superhuman accuracy for challenging image recognition tasks. However, the drive to improve accuracy often comes at a cost: modern state of the art networks require high computational resources beyond the capabilities of many mobile and embedded applications.

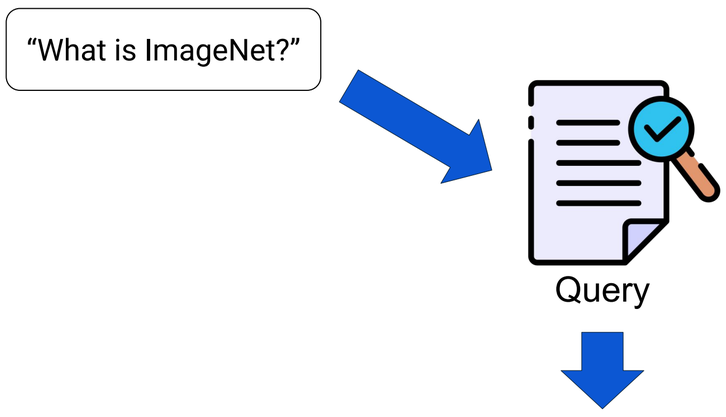
This paper introduces a new neural network architecture that is specifically tailored for mobile and resource constrained environments. Our network pushes the state of the art for mobile tailored computer vision models, by significantly decreasing the number of operations and memory needed while retaining the same accuracy.

Our main contribution is a novel layer module: the inverted residual with linear bottleneck. This module takes as an input a low-dimensional compressed representation which is first expanded to high dimension and filtered with a lightweight depthwise convolution. Finally, the official implementation is available as part of TensorFlow-Slim model library in [4].

This module can be efficiently implemented using standard operations in any modern framework and allows our models to beat state of the art along multiple performance points using standard benchmarks. Furthermore, this convolutional module is particularly suitable for mobile designs, because it allows to significantly reduce the memory footprint needed during inference by never fully materializing large intermediate tensors. This reduces the need for main memory access in many embedded hardware designs, that provide small amounts of very fast software controlled cache memory.

2. Related Work

Tuning deep neural architectures to strike an optimal balance between accuracy and performance has been an area of active research for the last several years. Both manual architecture search and improvements in training algorithms, carried out by numerous teams has lead to dramatic improvements over early designs such as AlexNet [5], VGGNet [6], GoogLeNet [7], and ResNet [8]. Recently there has been lots of progress in algorithmic architecture exploration included hyperparameter optimization [9, 10, 11] as well as various



PPX (search)

Verified Location

```
> ppx query "What is ImageNet?" --visualize
                                  Query Results
Query: What is ImageNet?
Results: 4 (min_score=0.5)
```

#	Score	Page	Element	BBox	Text
1	0.666	7	text	(654, 858, 1124, 1125)	For the ImageNet dataset, our archite...
2	0.615	8	reference…	(158, 723, 614, 820)	- [5] Alex Krizhevsky, Ilya Sutskever...
3	0.615	5	(aggregat…	(192, 341, 932, 1405)	# 6.1. ImageNet Classification
4	0.564	8	reference…	(158, 192, 614, 364)	- [1] Olga Russakovsky, Jia Deng, Hao...

4. References

[1] C. Cui et al., "PaddleOCR 3.0 technical report," arXiv:2507.05595, 2025.

[2] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," EMNLP-IJCNLP, 2019, pp. 3982–3992.

[3] Mistral AI, "Mistral OCR," 2025. mistral.ai/news/mistral-ocr

[4] Icons by Good Ware from Flaticon (flaticon.com)

5. Current Progress and Future Work

Jan — Feb — March — April

OCR & DocLayout

- PaddlePaddle OCR
- MistralOCR
- DocLayout

Algorithm & System Design

- Alignment algorithm
- Query using embeddings

Evaluation & App Implementation

- Create *ground truth* dataset
- Frontend app

Thesis Submission

- Draw conclusions
- Formulate findings